The project explores the application of the K-Means clustering algorithm to segment customers based on their behavioral and demographic data, offering actionable insights for businesses. The process begins with thorough data cleaning and analysis to ensure the dataset's reliability and usefulness. This involves identifying and addressing missing values, inconsistencies, and anomalies within the data. Both categorical and numerical data are analyzed to uncover patterns and distributions, with categorical attributes being encoded and numerical values standardized through Z-score normalization. Outliers are also handled carefully to prevent skewed results that could impact clustering accuracy.

Once the data is prepared, the K-Means algorithm is introduced as the core methodology for clustering. K-Means operates by iteratively assigning data points to clusters while minimizing the variance within each cluster and maximizing the differences between clusters. The project highlights the importance of initializing centroids, assigning points to the nearest cluster, updating centroids based on cluster means, and repeating these steps until convergence. This process ensures that the clusters are both cohesive and well-separated. The implementation of K-Means is further enhanced through parallelization techniques, such as leveraging Apache Spark. By distributing computations across multiple nodes, this approach ensures scalability and efficiency, making it well-suited for handling large customer datasets.

The evaluation of clustering results plays a critical role in the project, with the Sum of Squared Errors (SSE) metric being used to measure intra-cluster variance. By calculating SSE for various values of 𝐾, the project employs the Elbow Method to determine the optimal number of clusters. This ensures that the selected number of clusters balances compactness and simplicity. Additionally, the characteristics of each cluster are analyzed to uncover meaningful patterns and insights, providing a deeper understanding of customer behavior and preferences.

In conclusion, the project demonstrates the effectiveness of K-Means clustering for customer segmentation. By combining rigorous data preprocessing, algorithmic precision, and scalable computation, the approach delivers valuable insights that can inform targeted marketing strategies and enhance customer engagement. This methodology highlights the practical potential of machine learning in driving data-driven business decisions.